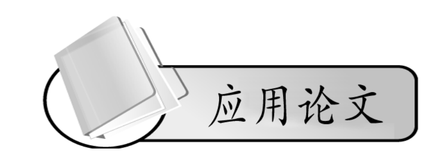


利用开放语义资源丰富个人名称规范数据
——基于 FOAF 的方案设计

郝嘉树
(国家图书馆　北京　100081)

摘要:【目的】我国规范数据质量差且维护效率低下，需探索低成本高效率的信息源获取模式，丰富个人名称规范数据。【方法】分别从语义资源数量和类型的有效性，高效维护的评价指标易获取、自动化程度、维护速度和开放资源可信度三个方面论证用语义资源维护名称规范的可行性，同时以 FOAF 为例设计实现方案。【结果】制定了获取语义资源的限制条件、接口方式和收割规则策略，给出发现、整合资源的 RDF 谓词以及开发包和软件两种实现技术，设计丰富名称规范数据的自动多重匹配算法和映射表。【局限】只提供实现流程及方案，没有付诸实施；语义资源获取后的存储方式、提取处理方法只是框架设计，没有给出详细的实现技术。【结论】可以将与个人相关的开放语义资源自动匹配，丰富本地名称规范数据。

关键词: 个人名称规范　开放语义资源　自动发现聚合　RDF　URI　FOAF　自动匹配丰富
分类号: G254　TP393.4

1　引　言

我国个人名称规范数据因不完整数据量大、重名严重及同名标目形式多样化等问题导致标目间区分度低，在维护规范数据和规范书目责任者时给编目员甄别和选择带来较大障碍，需要人工逐条分析、比对和辨别。编目员在个人名称规范维护和控制工作中花费的时间、精力与名称规范发挥的作用形成较大反差。

截至 2014 年底，国家图书馆名称规范数据量已高达 140 万条，7 年时间增长 62%，因此当前名称规范工作重点不应该是盲目扩张规范库规模，而是要靠高效获取和完善数据提升名称规范数据的维护效率和质量，从而真正发挥名称规范的区分、汇聚等功能。

探索低成本高效率的信息源获取模式是关键所在。目前，维护规范数据的信息来源有受编文献、工具书、期刊、网络、电话、邮件和交友软件等渠道，这些渠道大多是被动获取信息，可获得内容取决于是否提供和提供哪些信息；电话、邮件和交友软件虽然能得到指定信息，但沟通成本高不利于大规模开展。因此需探索新的信息源获取模式，在获得所需信息的同时，能快速大规模开展个人名称规范数据维护工作并降低人力物力成本。

国内外已有学者从开放关联数据的角度论述发布名称规范数据可以发现、整合信息，其中有些完成了数据的语义化发布，但并未给出如何实现规范数据的丰富和具体实施方案[1-2]；Elliott 提出利用 FOAF 帮助名称规范消歧，但只论述了潜在可能性，并提出相关疑问[3]。考虑到机构对自身数据不愿公开以及开放关联数据操作、管理的复杂性，本文提出不发布为开放关联数据，而是收割语义资源丰富本地名称规范的方法，并论述该方法可行的相关依据，同时以 FOAF 为例，制定语义资源获取策略、分析发现整合技术以及给出自动丰富名称规范数据的具体算法和方案。

通讯作者：郝嘉树，ORCID：0000-0002-4403-8516，E-mail：haojsh@nlc.cn。

2 开放语义资源：高效信息源获取模式

开放语义资源是发布在Web中以机器可理解可处理的资源描述框架(Resource Description Framework, RDF)模型和统一资源标识符(Uniform Resource Identifier, URI)表示的可分享、链接的数据集合[4]。语义数据通过“主体–谓词–客体”三元组(Triple)形式描述不同对象和它们之间的关系，资源用URI标识并用RDF模型表示后，经过发布，任何人都可以使用HTTP URI参引(Dereference，查找和获取)该数据。

开放语义资源是有效的名称规范信息源。目前互联网上发布的RDF三元组数量已是百亿级别，在这些开放语义资源中，包含了由网络用户发布的“自我申明”，它们通过用户创造内容(User Generated Content, UGC)[5]形式整合存在人头脑里有关人的事实信息，例如FOAF(Friend-Of-A-Friend)[6]、BibApp[7]和VIVO[8]，这些资源描述个人的兴趣爱好、开展的工作及项目、发表的著作及认识的朋友同事等，可用来补充、完善个人名称规范数据相关信息项；除此之外，开放语义资源中还有权威机构发布的名称规范档、人名表和叙词表等，如德国国家图书馆联合权威档(Gemeinsame Normdate, GND)[9]及国际虚拟规范文档(Virtual International Authority File, VIAF)[10]等基于RDF/XML描述的名称规范档，英国档案叙词表(UK Archival Thesaurus, UKAT)[11]及美国国会图书馆标题表(Library of Congress Subject Headings, LCSH)[12]等用简单知识组织系统(Simple Knowledge Organization System, SKOS)[13]表示的个人主题词，这些词汇表包含了权威的个人信息，其本身就可用来丰富名称规范附加成分、单纯参照和相关参照等，提高个人名称规范数据质量。

利用开放语义资源维护名称规范数据的高效性体现在易获取性、自动化程度和维护速度三个方面。目前国内外维护名称规范数据的主要方式是由编目员通过受编文献、工具书、期刊、网络和邮件等渠道查找责任者相关信息，并进行手工维护。利用开放语义资源的自动维护较之基于传统信息源的手工维护：

(1) 易获取性方面，前者的RDF三元组描述方式及URI技术容易在数据集之间跳转，将不同数据集合以各种关系形式连接起来，从而能极大程度上发现和获取资源，并且准确性高；而后者需人工逐一在各信息源中查找，并需辨识同名异人的情况。

(2) 自动化程度方面，前者结构化的数据可将RDF谓词和规范数据MARC字段建立映射，计算机程序能自动将语义资源收割到规范记录对应的字段中去；而后者需人工查找、辨识各字段对应的内容并手工输入。

(3) 维护速度方面，前者大部分流程都可以根据相关算法和RDF机制由机器自动、定向和批量地获取资源和维护规范数据，对于超出本地规范库范畴外的资源可快速新建记录扩张本地规范库规模；而后者在各环节主要靠人工来逐条比对、判断和维护，影响维护的速度。

从以上三个评价指标可以得出，利用开放语义资源的自动维护较之基于传统信息源的手工维护效率高，有利于个人名称规范数据质量及规模的提升。

开放语义资源可信度高。语义数据的技术架构提供可追踪来源的RDF语义描述方案，通过数据来源判断数据的可靠性，还有通过计算网络可信度推断来源名誉[14]；语义网为各类实体和所涉及的大量概念、术语提供了规范控制，使得提及某一实体或概念术语时，系统能自动给予归并或参照，这种规范控制的结果就使信息在一定程度上更加可信[15];开放的语义资源中，网络用户发布的“自我申明”是个人对自我的真实反映，排除恶意欺诈，该种模式下申明的内容是客观的；发布的规范文档和叙词表等是由权威机构编制，内容准确真实。

调查发现FOAF是最受欢迎的本体之一[16]，为方便理解，本文即以FOAF为例，设计用语义资源丰富个人名称规范数据的方案，探究如何识别和获取开放语义资源进行个人信息的自动发现及聚合，并自动匹配和丰富名称规范数据。

3 FOAF

FOAF是网络用户用已定义好的RDF词汇表形式化描述个人信息和其相关的社会网络，其本质为描述个人的简单本体。它由Brickley等于2000年创建，遵循W3C体系，最初只描述个人，后扩展到各类群体，如机构、公司和地点，FOAF描述词汇随之历经10次更新，于2014年最终确定下来[17]。

计算机对FOAF文档可读可理解，文档经发布便可搜索和处理。FOAF命名空间用RDF Schema定义

的词汇(标签)描述个人及相关属性(信息项), 计算机通过这些标签理解和处理FOAF文档; FOAF文档可使用“FOAF-a-Matic 2.0”[18]、“Quatuo”[19]等在线工具生成, 也可参考 RDF/XML 语法和 FOAF 词汇手工创建。获取有关人的信息并不容易, FOAF 通过 UGC 形式整合存在人头脑里有关人的事实信息, 并通过 URI 分散的协同形成社会网络。本文所使用的 foaf:Person(个人)类包含的属性如表 1 所示:

表 1 foaf: Person 包含的属性标签及说明

属性标签	说明
foaf:pastProject	曾做过的项目
foaf:currentProject	正在进行的项目
foaf:publications	发表的文章
foaf:knows	认识的人
foaf:workplaceHomepage	工作单位网站
foaf:workInfoHomepage	从事工作的相关信息
foaf:schoolHomepage	学校信息
foaf:firstName	名
foaf:surname	姓氏
foaf:lastName	姓
foaf:gender	性别
foaf:birthday	出生日期
foaf:mbox	邮箱 URI
foaf:mbox_sha1sum	加密的邮箱 URI
foaf:interest	兴趣
foaf:geekcode	网络个性签名
foaf:img	图片
foaf:plan	工作、个人生活计划
foaf:jabberID、foaf: aimChatID、foaf:yahooChatID、foaf:icqChatID	各种网络账号
foaf:myersBriggs	迈尔斯布里格斯类型指标(表征人的性格)

其中, 用 foaf:mbox 或 foaf:mbox_sha1sum 作为识别个人的 URI, 名字等不具唯一标识性, 而不同人使用不同的邮箱, FOAF 用邮箱代表背后使用的人[20]。foaf:knows 表示认识的人, 通过该标签可以很容易把相关人员和实体关联起来形成社会网络, 从而丰富个人规范数据相关参照。

4 基于 FOAF 丰富个人名称规范数据的方案设计

利用 FOAF 丰富名称规范数据流程如图 1 所示。

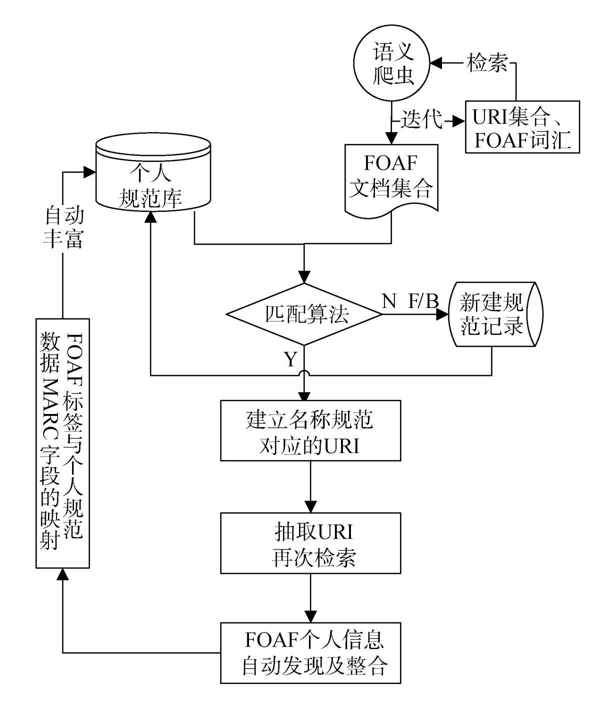


图 1 利用 FOAF 丰富个人名称规范数据的方案设计

使用语义爬虫软件根据限制条件获取 FOAF 文档集合, 采用相关算法和名称规范数据进行匹配, 匹配成功则利用 URI 自动发现聚合个人信息, 并根据 FOAF 标签与规范数据 MARC 字段的映射自动完善规范数据, 匹配失败则对 FOAF 文档进行筛选用于名称规范数据的新建, 从而扩大规范库种类和数量, 具体实施方案如下。

4.1 识别和获取 FOAF 文档

选用开放语义资源作为信息源的一个重要原因是可由计算机自动批处理大量数据, 而人只是制定规则。在获取资源之前, 要制定限制条件、类型和规则, 保证获取资源的有效性, 具体方案如下:

(1) 识别限制条件

数据集满足下面条件即视为 FOAF 文档:

①是有效的 RDF 文档;

②文档使用 FOAF 命名空间;

③X 是 rdf:Property 的实例并且来自 FOAF 命名空间;

④在主体位置只能有一个类型为 foaf:Person 的实例并且不能作为文档中任何三元组的客体。另外,文档中出现的其他 foaf:Person 实例必须作为客体, 不能出现在主体位置。

其中涉及的 FOAF RDF 模型如图 2[16]所示。

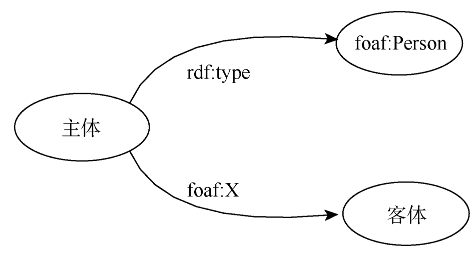


图 2　FOAF 文档模型

在具体实施时, RDF 的有效性使用 RDF 解析器验证; 命名空间通过检索是否包含"http://xmlns.com/foaf/0.1/"判断; 条件③可以具化为判断 foaf:X 的 rdfs:domain(定义域)是否为 foaf:Person; 条件④是理想的 FOAF 文档, 在实际中过于严苛, 可以将其简化, 排除没有揭示实例的情况就是符合定义的 FOAF 文档[16]。

(2) 获取接口的选择

关联数据的获取方式有 5 类: SPARQL 查询准确率高, 但需掌握相关语法及获取的资源少; Web Service API 可长期、批量获取数据, 但需熟悉 API 背后的各种协议; 批量下载简单直接, 但不适用于大规模、更新频率高的数据源; 动态网页抽取和语义搜索引擎/爬虫能同时获取多个数据源资源, 但受自身算法、策略影响大[21]。结合以上 5 种获取方式的优势和劣势, 本文选用语义爬虫方式, 如 LDSpider[22], 目的是发现更多未知、可能存在的语义数据集。

(3) 收割规则

为不引起版权纠纷, 在搜索数据集时要注意数据规定的访问权限, 选择访问方式为免费和可开放获取的关联数据集; 对于有限制的数据集获取其公开部分的元数据内容。

获取 FOAF 文档是迭代的过程。将前述规则转化为语义爬虫中有效提问, 相继搜索包含 FOAF 命名空间、foaf:Person 及文件类型为.foaf 的文档和后缀为.rdf、.xrdf、.owl 的 RDF 文档, 从而收集 FOAF 文档和 URIs 集合。并再次在语义爬虫中利用已搜集的 URIs 和 FOAF 词汇, 如 foaf:knows 和 rdfs:seeAlso, 发现新的 FOAF 文档。通过在爬虫中搜索发现, FOAF 文档主要来自于博客, 常用标签排名为 foaf:mbox_sha1sum、foaf:nick、foaf:name、foaf:homepage、foaf:knows、foaf:birthday、foaf:interest 和 rdfs:seeAlso 等[16]。

4.2　FOAF 个人信息的自动发现及聚合

RDF 具有开放性和互联性, 实体经 RDF 描述、发布后可被计算机检索和处理, 并可将网络上离散的数据片段自动关联起来发现新内容。数据之间的链接主要依靠三元组中谓词和 URI 的使用, 其中谓词只要根据应用领域选择相应的 RDF 属性即可[23]。

如在 FOAF 中, 从个人发布的 FOAF 文档中获取信息, 还通过唯一标识 foaf:mbox 或 foaf:mbox_sha1sum 或 foaf:homepage 游历到另外的 FOAF 文档, 从而自动发现和整合文档集合中有关此人的所有信息, 也因此可用来完善规范数据的附加成分、单纯参照和注释, 提高个人名称规范数据质量; FOAF 通过 foaf:knows 将自己的朋友、同事或认识的人关联起来, 再通过唯一标识匹配 FOAF 文档集中不同个体的 owl:sameIndividualAs(个体相同)关系[24], 这样分散的文档集合就能形成社会网络, 可帮助构建、完善个人名称规范数据的相关参照。

通过分析语义数据集的谓词和获取 URI 集合, 就可以使用开发包或已有的语义搜索软件发现和整合语义资源。其中, 对 RDF 数据有处理能力的开发包有 Jena[25]、Sesame[26]和 PHP[27]等, 已有的语义搜索软件有 Sindice[28]、Swoogle[29]等。图 3 是利用 FOAF 搜索器 NetEstate[30]对上海图书馆刘炜自动发现和整合的结果。



图 3　FOAF 资源的自动发现及整合

通过 URI (foaf_sha1sum)排除同名异人, 聚合同人异名, 检索结果都为上海图书馆刘炜的相关信息; 通过相关谓词, 即 FOAF 标签自动发现、聚合同一信息项, 如"刘炜"本人发布的 FOAF 文档姓名为"Keven Liu", 通过 foaf_sha1sum 及 foaf:name 可自动发现文档集合中他人描述"刘炜"使用的名字"Keven Liu"和"刘炜", 并通过 foaf:name 标签整合显示; 另外 foaf:knows 将文档集中所有认识此人和此人认识的人自动聚合在一起, 即用嵌套在 foaf:knows 中描述的结构进行关联,

最终形成个人社会网络。

4.3　FOAF 与名称规范数据的自动匹配及丰富

为高效开展规范数据的丰富，需开发自动匹配算法识别规范记录对应的 FOAF 唯一标识。“姓名/生卒年/著作”、“姓名/生卒年”和“姓名/著作”组合对个人的识别度依次降低，可根据信息完整程度和是否匹配成功，逐一采用这三种组合进行自动识别；另外姓名作为匹配的主要内容，要充分利用名称规范数据的变异名称和 FOAF 昵称，使同一人的不同名称形式都参与比较提高匹配率。规范数据匹配 FOAF 唯一标识的算法如下：

(1) 分别抽取名称规范记录 200 字段名称及其 400 字段变异形式、200$f 和挂接书目数据 200$a$e，建立集合 N_{ij}、B_i 和 W_i(N_{ij} 为规范记录 i 的第 j 个名称，B_i 和 W_i 分别为规范记录 i 的生卒年和著作)；

(2) 分别抽取 FOAF 文档 foaf:lastName+foaf:firstName 及 foaf:nick、foaf:birthday 和 foaf: publications 建立集合 F_{mn}、B_m 和 W_m(F_{mn} 为 FOAF 文档 m 的第 n 个名称，B_m 和 W_m 分别为 FOAF 文档 m 的生卒年和著作)；

(3) 将“F_{nm}/B_n/W_n”与“N_{ij}/B_i/W_i”匹配，即采用“姓名/生卒年/著作”模式并将名称的各种变异形式都与生卒年、著作组合进行比对；

(4) 根据匹配结果再逐一用“姓名/生卒年”和“姓名/著作”进行比对；

(5) 识别成功的建立规范记录对应的 FOAF 唯一标识(foaf:mbox 或 foaf:mbox_sha1sum 或 foaf:homepage)，以便名称规范数据的完善和定期抓取、更新数据；

(6) 未识别成功的，筛选 “姓名/生卒年” 模式中生卒年信息完整的 FOAF 文档作为新记录，用来丰富名称规范数据种类和数量(著作是否完整无法判断，因此不包括另外两种模式)。

另外，在匹配过程中要对数据进行相关处理、算法进行调整以提高匹配率：

(1) 对于姓名，中、日、韩名称直接抽取规范数据 200 和 400 字段的$a 与 FOAF 中 foaf:lastName+foaf:firstName、foaf:nick 匹配；而外国人的拉丁文名称则选取 200 字段$c(去除其中逗号)、400 字段$a+$b(去除之间逗号)与 foaf:lastName+foaf:firstName、foaf:nick 匹配，如果比对未成功，则对姓名进行缩写以提高匹配率。

(2) 对于生卒年，有些规范数据因著录错误将生卒年著入$a 和$c 中，可通过判断是否为日期数据获取该信息；对于完整生卒年信息未匹配成功的情况，可逐一去除生卒年末尾数、卒年或生年以提高匹配率[31]。

(3) 对于题名，对书目数据中不规范、错误著录进行处理，如检验是否为题名和删除当中的姓名；去除两类集合中题名的标点符，如破折号、方括号、冒号等；对于拉丁文要去掉前后空格、去除开头冠词和助词和不区分大小写等[32]。

计算机可以解析 RDF 数据含义，通过语义标签定向找到相关信息，因此只要将个人 RDF 资源语义标签和规范记录 MARC 字段建立映射，计算机程序就能自动将 RDF 数据收割到规范记录对应的字段中去，用于完善名称规范数据。在 FOAF 中，根据名称规范记录揭示的个人信息项寻找与之对应的 FOAF 属性，并将属性对应的标签和名称规范 MARC 字段及子字段建立映射，如表 2 所示：

表 2　FOAF 标签与个人规范数据 CNMARC 字段的映射

CNMARC	字段解释	FOAF 词汇	说明	重复与否
091$aFOAF $bURI	开放数据类型 FOAF 及 URI	foaf:mbox 或 foaf:mbox_sha1sum	方便数据定期维护	可重复
091$aSKOS $bURI	与 FOAF 对应的 SKOS 及 URI	foaf:focus	与 SKOS 搭配使用，帮助指明不同 SKOS 体系中的个人和团体	可重复
120$a	编码数据字段	foaf:gender	区分于 200$c 职业行业	唯一
200$c	附加成分	foaf:interest	职业、行业	可重复
200$f	生卒年	foaf:birthday		唯一
391$a	发表著作	foaf:publications		可重复
391$b	开展项目	foaf:pastProject、foaf:currentProject		可重复
391$c	工作计划	foaf:plan		可重复

(续表)

CNMARC	字段解释	FOAF 词汇	说明	重复与否
392$a	性格	foaf:myersBriggs、foaf:geekcode		可重复
392$b	博客	foaf:weblog		可重复
392$c	人物肖像	foaf:image	指向图片库	可重复
393$a	工作单位	foaf:workInfoHomepage、foaf:workplaceHomepage		可重复
393$b	学校	foaf:schoolHomepage		可重复
400$a	单纯参照	foaf:name 或 foaf:lastName+ foaf:firstName、foaf:nick、foaf:yahooChatID、foaf:skypeID、foaf:icqChatID	其他形式的名字、昵称及网络账号	可重复
500$a	相关参照	foaf:knows	相关的人与机构	可重复
810$a	参考数据源	URI	发布的 URI 地址	可重复

启用和扩展新字段对个人名称规范数据中信息进行结构化处理。我国名称规范格式中, 200 字段附加成分$c、300 字段个人相关信息并没有进行区分, 为符合当下编目主流趋势, 适应 RDA 规则及新修订的 UNIMARC 规范格式, 也方便名称规范库后续开发利用, 建议启用和扩展新字段对个人信息进行结构化处理。其中, 启用 120 字段用于区分 200 字段附加成分性别与职业; 因 FOAF 多个属性与 300$a 对应, 新增 391、392、393 字段分别著录个人工作科研情况、兴趣性格和相关团体信息, 其中包含的子字段揭示更具体的信息项; 新增 091 字段记录对应语义数据的唯一标识, 开放数据处于动态变化中, 通过唯一标识可定期完善数据。

5 结 语

本文针对名称规范数据质量差且维护效率低下的情况, 提出收割开放语义资源丰富本地名称规范数据的方法, 在论述其可行性的基础上以 FOAF 为例设计实现流程, 并制定识别获取语义资源的限制条件、接口方式和收割规则策略, 给出发现、整合资源的 RDF 谓词以及两种实现技术开发包和软件, 设计自动丰富名称规范数据的多重匹配算法和映射表。FOAF 只是开放语义资源的一个典型应用, 只要是与个人相关的开放语义资源, 如 VIAF、GN、CNO(CSHL Name Ontology, 冷泉港实验室姓名本体)和 SKOS 表示的个人主题数据等, 都可用于自动发现和收割语义信息, 丰富本地个人名称规范数据。

下一步将开展试验工作, 重点关注匹配算法的效果, 根据结果调整匹配策略和相关参数; 另外语义资源与一般资源不同, 带有语法和格式, 因此要研究数据存储方式和提取、处理的实现技术。

参考文献:

[1] Myntti J, Cothran N. Authority Control in a Digital Repository: Preparing for Linked Data [J]. Journal of Library Metadata, 2013, 13(2-3): 95-113.

[2] 刘炜, 张春景, 夏翠娟. 万维网时代的规范控制[J]. 中国图书馆学报, 2015, 41(3): 22-33. (Liu Wei, Zhang Chunjing, Xia Cuijuan. Authority Control for the Web [J]. Journal of Library Science in China, 2015, 41(3): 22-33.)

[3] Elliott S. Survey of Author Name Disambiguation: 2004 to 2010 [J/OL]. Library Philosophy & Practice. http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1459&context=libphilprac.

[4] Schreiber G, Raimond Y. PDF 1.1 Primer [EB/OL]. (2014-06-24). [2015-06-07]. http://www.w3.org/TR/rdf11-primer/.

[5] UGC [EB/OL]. [2015-10-21]. http://baike.baidu.com/subview/713949/9961909.htm.

[6] FOAF (2000-2015+) [EB/OL]. [2015-08-07]. http://www.foaf-project.org/.

[7] BibApp [EB/OL]. [2015-10-21]. http://bibapp.org/.

[8] VIVO [EB/OL]. [2015-10-10]. http://www.vivoweb.org/.

[9] Gemeinsame Normdate [EB/OL]. [2015-06-15]. http://d-nb.info/standards/elementset/gnd.

[10] Virtual International Authority File [EB/OL]. [2015-06-15]. http://www.viaf.org/.

[11] UK Archival Thesaurus [EB/OL]. [2015-06-15]. http://www.ukat.org.uk/.

[12] Library of Congress Subject Headings [EB/OL]. [2015-06-15]. http://id.loc.gov/authorities/subjects.html.

[13] Simple Knowledge Organization System [EB/OL]. (2012-12-13). [2015-06-15]. http://www.w3.org/2004/02/skos/.

[14] Golbeck J, Parsia B. Trust Network-Based Filtering of Aggregated Claims [J]. International Journal of Metadata, Semantic and Ontologies, 2006, 1(1): 58-65.

[15] 刘炜. 关联数据：概念、技术及应用展望[J]. 大学图书馆学报, 2011, 29(2): 5-12. (Liu Wei. Overview on Linked Data: Concept, Technology and Implementation [J]. Journal of Academic Libraries, 2011, 29(2): 5-12.)

[16] Ding L, Zhou L, Finin T, et al. How the Semantic Web is Being Used: An Analysis of FOAF Documents [C]. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences. IEEE, 2005: 113-121.

[17] FOAF Vocabulary Specification 0.99[EB/OL]. [2015-07-11]. http://xmlns.com/foaf/spec/.

[18] FOAF-a-Matic [EB/OL]. [2015-07-31]. http://www.ldodds.com/foaf/foaf-a-matic.en.html.

[19] Quatuo [EB/OL]. [2015-07-31]. http://www.quatuo.com/.

[20] Dumbill E. XML 观察：使用 XML 和 RDF 找到朋友[EB/OL]. [2015-08-18]. http://www.ibm.com/developerworks/cn/xml/x-watch/part3/index.html. (Dumbill E. Finding Friends with XML and RDF [EB/OL]. [2015-08-18]. http://www.ibm.com/developerworks/cn/xml/x-watch/part3/index.html.)

[21] 王思丽，马建玲，李慧佳，等. 关联数据集中开放资源的自动获取研究[J]. 图书馆学研究, 2015(18): 49-54. (Wang Sili, Ma Jianling, Li Huijia, et al. Research on Obtaining Open Resource in the Linked Data Automatically [J]. Research on Library Science, 2015(18): 49-54.)

[22] LDSpider [EB/OL]. [2015-07-31]. https://github.com/ldspider/ldspider.

[23] Graves M, Constabaris A, Brickeley D. FOAF: Connecting People on the Semantic Web [J]. Cataloging & Classification Quarterly, 2009, 43(3-4): 191-202.

[24] Motik B, Parsia B. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (2nd Edition) [EB/OL]. (2012-12-11). [2015-08-12]. http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/#Symmetric_Object_Properties.

[25] Apache Jena [EB/OL]. [2015-10-21]. http://jena.apache.org/.

[26] Sesame [EB/OL]. [2015-10-21]. http://rdf4j.org/.

[27] PHP [EB/OL]. [2015-10-21]. http://php.net/.

[28] Sindice: Disambiguation Page [EB/OL]. [2015-10-21]. http://www.sindice.com/.

[29] Semantic Web Search-Swoogle [EB/OL]. [2015-10-21]. http://swoogle.umbc.edu/.

[30] NetEstate. Friend of A Friend (FOAF) Search Engine [EB/OL]. [2015-10-21]. http://www.foaf-search.net/.

[31] FRBR Work-Set Algorithm Version 2.0[EB/OL]. [2015-10-11]. http://www.oclc.org/content/dam/research/activities/frbr-algorithm/2009-08.pdf.

[32] FRBR Work-Set Algorithm [EB/OL]. [2015-10-11]. http://www.oclc.org/content/dam/research/activities/frbralgorithm/2005-04.pdf.

利益冲突声明：

作者声明不存在利益冲突关系。

收稿日期: 2015-08-23
收修改稿日期: 2015-10-23

Enriching Personal Name Authority with Open Semantic Resources: FOAF for Schema Design

Hao Jiashu
(National Library of China, Beijing 100081, China)

Abstract: [**Objective**] This study ctreated a new model to improve the quality and maintenance efficiency of the personal name authority data in China. [**Methods**] To prove the feasibility of using open semantic resources to enrich the name authority data, this study analyzed the number and types of semantic resources, evaluation metrics, automation and maintenance speed, as well as the credibility of the open resource. The FOAF was used as an example to implement the schema. [**Results**] This study set restriction conditions, interface mode and harvest rules for obtaining the semantic resources. It created RDF predicate and two realizing techniques, like SDK and software to discover and integrate resources. This study designed automatic multi-matching algorithm and mapping table to automatically enrich name authority data. [**Limitations**] Only creates the schema, which was not put into practice. The semantic resource's storage model and the extraction processing methods are also at the initial framework stage. No detailed implementation technology was discussed. [**Conclusions**] The proposed method could be automatically matched with open semantic resources of the individual names to enrich local personal name authority data.

Keywords: Personal name authority　Open semantic resources　Automatic discovering and aggregating　RDF　URI　FOAF　Automatic matching and enriching

---

耶鲁大学图书馆加入人文科学开放图书馆合作资助系统

耶鲁大学图书馆加入了人文科学开放图书馆(OLH)的图书馆合作资助系统。此次合作将为耶鲁大学和世界各地的学者提供数量更多、质量更高的人文科学开放获取出版物。

OLH 的创始人和学术项目主管 Martin Paul Eve 指出: “我很高兴耶鲁大学能够加入我们。很显然, 每个人都将受益于开放研究, 但我们必须找到一种方法促进人文学科发展符合经济规律。在耶鲁这样的机构帮助下, 我们将实现这一目标。” 人文科学开放图书馆是一个学术主导、无需作者支付费用的金色开放获取出版平台。该平台的启动资金来自于 Andrew W. Mellon 基金会, 并且该平台的其他费用均由国际化的图书馆联盟支付, 而不收取任何形式的作者费用。

耶鲁大学馆藏建设主任 Daniel Dollar 补充说: “学术交流新的资助模式仍然是一个重要的试验领域。耶鲁大学图书馆很高兴能够支持学术资源的公共获取, 正如人文科学开放图书馆所预想的那样。” 在这种情况下, Arcadia 基金的慷慨资助使得本次合作成为了可能。

(编译自: http://web.library. yale.edu/news/2015/06/yale-university-library-joins-open-library-humanities-library)

(本刊讯)